ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ТЕЛЕКОММУНИКАЦИИ
INFORMATION TECHNOLOGY AND TELECOMMUNICATIONS

УДК004.056.52

DOI: 10.21822/2073-6185-2024-51-2-77-82 Оригинальная статья/ Original article

Модуль голосовой аутентификации с использованием мел-кепстральных коэффициентов

Д.А. Елизаров, П.А. Ашаева, Е.А. Степанова
Омский государственный университет путей сообщения,
644046, г. Омск, пр. Маркса, 35, Россия

Резюме. Цель. Целью исследования является разработка и применение способа извлечения информации о личности пользователей из записей их голоса при помощи вычисления мел-кепстральных коэффициентов. **Метод.** В ходе исследования использованы методы извлечения информативных признаков из голосовой записи, позволяющие идентифицировать диктора, а также представлена схема аутентификации с использованием мел-кепстральных коэффициентов. **Результат.** На основе данного метода был реализован модуль аутентификации с использованием аудио записей голоса пользователей простейшими MFCC. Модуль аутентификации был разработан с использованием языка Python. **Вывод.** Метод биометрической аутентификации является недорогим и относительно простым способом проверки подлинности пользователей. Несмотря на очевидные преимущества мел-кепстральных коэффициентов, данный метод имеет определенные недостатки, для устранения которых могут быть использованы различные частотные фильтры, а также сторонние алгоритмы анализа аудиозаписей.

Ключевые слова: биометрия, голосовая аутентификация, мел-кепстральные коэффициенты

Для цитирования: Д.А. Елизаров, П.А. Ашаева, Е.А. Степанова. Модуль голосовой аутентификации с использованием мел-кепстральных коэффициентов. Вестник Дагестанского государственного технического университета. Технические науки. 2024; 51(2):77-82. DOI:10.21822/2073-6185-2024-51-2-77-82

Voice authentication module using mel-cepstral coefficients

D.A. Elizarov, P.A. Ashaeva, E.A. Stepanova
Omsk State Transport University,
35 Marx Ave., Omsk 644046, Russia

Abstract. Objective. The purpose of the study is to develop and apply a method for extracting information about the identity of users from recordings of their voices using the calculation of mel-cepstral coefficients. **Method.** In the study of the application of methods for extracting informative features from a voice recording, allowing identification of the speaker, an authentication scheme using mel-cepstral coefficients is presented. **Result.** Based on this method, an authentication module was implemented using audio recordings of user voices using the simplest MFCC. The authentication module was developed using Python language **Conclusion.** The biometric authentication method is an inexpensive and relatively simple way to verify the authenticity of users. Despite the obvious advantages of mel-cepstral coefficients, this method has certain disadvantages. To eliminate shortcomings, various frequency filters can be used, as well as third-party algorithms for analyzing audio recordings.

Keywords: biometrics, voice authentication, mel-cepstral coefficients

For citation: D.A. Elizarov, P.A. Ashaeva, E.A. Stepanova.Voice authentication module using mel-cepstral coefficients. Herald of Daghestan State Technical University. Technical Sciences. 2024; 51(2):77-82. DOI:10.21822/2073-6185-2024-51-2-77-82

Введение. Процесс аутентификации представляет собой комплекс мер, позволяющих определить правомерность доступа пользователя к некоторым ресурсам посредством предоставления им аутентифицирующего фактора [1–2]. В настоящее время существуют различные способы аутентификации, одним из которых является биометрическая аутентификация. При биометрической аутентификации в качестве признака пользователь должен предоставить физиологический или поведенческий признак. Одним из таких признаков может быть голос. Биометрия позволяет упростить процесс аутентификации для пользователя, но в то же время для её проведения необходимо использовать эффективное оборудования и методы верификации, чтобы исключить возможность нелегитимного доступа или отказа в доступе разрешённым пользователям.

Постановка задачи. Голосовая аутентификация может стать недорогим простым способом проверки подлинности пользователей, в связи с чем необходимо исследовать характеристики голосовых шаблонов и разрабатывать методы обработки таких шаблонов. В данной работе рассмотрены основные методы извлечения информативных признаков из голосовой записи, позволяющих идентифицировать диктора, а также представлена схема аутентификации с использованием мел-кепстральных коэффициентов.

Ожидается, что объем рынка голосовой биометрии вырастет с 1,93 млрд долларов США в 2023 году до 4,18 млрд долларов США к 2028 году при среднегодовом темпе роста 16.75% в течение прогнозируемого периода (2023-2028 годы) [2]. В связи с этим все более актуальной становится проблема повышения безопасности систем голосовой аутентификации.

Методы исследования. Существуют различные классификации биометрических параметров, которые могут быть использованы для аутентификации. Часто биометрические параметры делят на две группы: физиологические и поведенческие.

Физиологические параметры также называют статическими, поскольку в нормальном случае данные характеристики однозначны, неизменяемы и не зависящие от внешних факторов. Однако, данные характеристики могут изменяться, например, в случае серьёзной болезни или несчастного случая.

Поведенческие, или динамические, параметры могут изменяться в зависимости от эмоционального или физического состояния человека. Голос как параметр аутентификации является именно поведенческой характеристикой, поскольку зависит от физического состояния голосового тракта. Свойства голоса (частота, интонация, носовой звук и др.) – это уникальные характеристики, описывающие строение речевого тракта говорящего [3].

Мел-кепстральные коэффициенты. Аутентификация не может быть проведена непосредственно по записи голоса диктора, так как запись в амплитудной области отражает громкость, произнесённые слова и другие параметры, не относящиеся к характеристикам голоса диктора. Поэтому для выделения характеристик самого голоса необходимо провести обработку записей. В ходе обработки записи выполняется вычисление определённого числа коэффициентов, отражающих совокупность признаков голоса. Среди таких коэффициентов наиболее популярными являются мел-кепстральные коэффициенты (от англ. Mel Frequency Cepstral Coefficients, MFCC).

Мел определяет высоту звука – восприимчивость к изменению частоты звукового сигнала слуховым аппаратом человека [4]. Данную величину можно получить по формуле 1, полученной после проведения ряда психофизических экспериментов:

$$m = 2595 * \log_{10}(1 + \frac{f}{700}) \quad (1)$$

где m – высота звука (мел), f – частота звука (Гц).

Спектр является представлением исходного сигнала в частотной области [5]. Такое преобразование позволяет снизить информационную избыточность: характеристики сигнала представляются в более компактном виде по сравнению с его временной формой. Данное утверждение справедливо для простейших сигналов, например, синусоидальных.

На практике чаще всего встречаются сложные сигналы, являющиеся совокупностью большого количества более простых сигналов. В связи с этим, для них спектральное представление также может включать большое количество информации, затрудняющее их дальнейшую обработку. На основе этого было введено новое понятие – кепстр.

Кепстр является обратным преобразованием Фурье от логарифма спектрального представления сигнала, т. е. кепстр – это спектр логарифма спектра временной волны [6]. Кепстр позволяет представить сигнал в ещё более усечённом виде (без потери информативности).

Процедура вычисления мел-кепстральных коэффициентов включает в себя следующие действия:

1) усиление высоких частот;

2) сегментирование сигнала на фреймы;

3) применение оконной функции и выполнение быстрого дискретного преобразования Фурье;

4) мел-частотное преобразование;

5) выполнение дискретного косинусного преобразования.

Усиление высоких частот исходного сигнала не обязательное, но, в некоторых случаях, желаемое действие, поскольку форманты высоких частот при обработке сигнала имеют меньшую амплитуду, по сравнению с более низкими частотами [7]. Данное действие может быть выполнено по формуле 2.

$$y'(t) = y(t) - a * y(t-1) \tag{2}$$

где $y'(t)$ – сигнал с усиленными высокими частотами, $y(t)$ – исходный дискретизированный сигнал в момент времени t, a – коэффициент усиления.

Коэффициент a принимает значение от 0,95 до 0,98 включительно.

Сегментирование сигнала на фреймы необходимо для дальнейшего упрощения вычислений, операции выполняются к каждому фрейму, независимо от других. Однако, для сохранения зависимости сегментов, фреймы могут пересекаться.

Применение оконной функции и выполнение быстрого дискретного преобразования Фурье к каждому полученному фрейму условно можно объединить в один шаг. Оконные функции используются для для улучшения спектрального разрешения сигнала [6]. Оконное преобразование Фурье представлено 3.

$$F(m,\omega) = \sum_{n=0}^{N-1} f(n) * W(n-m) * e^{-\frac{2\pi i}{N} * n} \tag{3}$$

где $W(n-m)$ – оконная функция, N – количество отчётов, $f(n)$ – значение сигнала в точке n, n индекс частоты [8].

В качестве оконной функции для вычисления MFCC наиболее часто используют функцию Хэмминга (формула 4).

$$W(n) = 0.54 - 0.46 * \cos(\frac{2 * \pi * n}{N-1}) \tag{4}$$

В зависимости от типа задач, помимо функции Хэмминга можно также использовать прямоугольную, Хеннинга, Блэкмана и другие [9]. После преобразования Фурье сигнал в каждом фрейме представлен в амплитудно-частотной области. Теперь необходимо выполнить мел-фильтрацию. Преобразование частот в мел-шкалу подразумевает выбор числа мел-полос пропускания и вычисление мел-фильтра полос пропускания. Число мел-полос является гиперпараметром, который подбирается эмпирически в зависимости от природы звука.

Для вычисления мел-фильтра нужно:

1) перевести нижнюю и верхнюю частоты сигнала в мелы (1);

2) полученный отрезок разделить на полосы;

3) перевести крайние точки каждой полосы перевести из мелов в Гц (формула 5);

4) построить треугольные фильтры.

$$f = 700 * (10^{m/2595} - 1) \quad (5)$$

Дискретное косинусное преобразование является заменой обратному преобразованию Фурье. Использование DCT обусловлено его простотой, а также возможностью получить вещественные MFCC, а не комплексные (формула 6).

$$\hat{C}_n = \sum_{n=1}^{k} (\log \hat{S}_k) * \cos[n * (k - \frac{1}{2}) * \frac{\pi}{k}] \quad (6)$$

где $\hat{C}_n$ – n-ый мел-кепстральный коэффициент, k – число мел-кепстральных коэффициентов, $\hat{S}_k$ – результат мел-фильтрования (энергетический спектр каждого фрейма).

На основе экспериментов, было вычислено оптимальное число коэффициентов, равное 12 – 13. В первых коэффициентах содержится больше всего полезной информации. Помимо этого, на практике часто вычисляют разницу этих коэффициентов (дельты).

У MFCC, тем не менее, есть ряд недостатков, одним из которых является неустойчивость к шуму [10]. Если при записи голоса пользователя на фоне имеется шум (другие голоса, звуки, помехи микрофона), то коэффициенты сильно изменятся, т.е. в них будет информация не только о голосе, но и о шуме. Для устранения этого недостатка на практике часто используют фильтры высоких и низких частот.

Обсуждение результатов. Согласно выше описанному алгоритму была разработана модель аутентификации, использующая простейшие MFCC. Процесс аутентификации проходит следующим образом:

1) записывание голоса диктора, произносящего псевдо-случайно сгенерированную фразу;

2) выполнение амплитудной нормировки сигнала;

3) разбиение записи на пересекающиеся на 10 миллисекунд фреймы;

4) получение спектра сигнала и вычисление периодограммы каждого фрейма;

5) вычисление общей энергии сигнала каждого фрейма;

6) получение блока мел-фильтров;

7) выполнение скалярного произведения мел-фильтров и периодограммы каждого фрейма;

8) логарифмирование полученной энергии каждого фильтра;

9) выполнение дискретного косинусного преобразования;

10) сравнение полученных коэффициентов с эталонными с помощью алгоритма динамической трансформации временной шкалы и вычисления евклидового расстояния.

Модуль аутентификации был разработан с использованием языка Python [11 – 15]. Перед выполнением аутентификации необходимо получить эталонные MFCC. Чтобы повысить надёжность системы, нужно сделать несколько записей. В данной работе для получения эталонных MFCC выполняется 10 записей голоса диктора, при этом каждый раз пользователю необходимо произнести сгенерированную фразу (рис. 1).

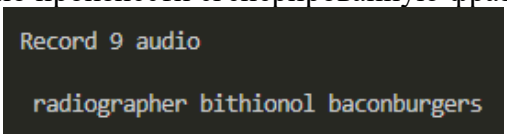

Рис. 1. Запись голоса для вычисления эталонных коэффициентов
Fig. 1. Voice recording for calculating reference coefficients

Процесс аутентификации, как описано выше, происходит аналогично: пользователю генерируется фраза, которую нужно произнести. После этого, вычисленные коэффициенты сравниваются с эталонными, т.е. происходит сравнение с коэффициентами для каждой эталонной записи (рис. 2).

```
 spackled ungifted sannyasi

End of recording

Authenticating.... Please wait.....
D:/Голосовая аутентификация/test/Polina_source_mfcc_coefficients_0.txt
dist = 23.559704126408974   threshold = 27   avdist = 23.559704126408974

D:/Голосовая аутентификация/test/Polina_source_mfcc_coefficients_1.txt
dist = 23.595703978618513   threshold = 27   avdist = 23.577704052513745

D:/Голосовая аутентификация/test/Polina_source_mfcc_coefficients_2.txt
dist = 24.489918627387674   threshold = 27   avdist = 23.881775577471718

D:/Голосовая аутентификация/test/Polina_source_mfcc_coefficients_3.txt
dist = 22.61091987405287   threshold = 27   avdist = 23.564061651617006
```

Рис. 2. Сравнение вычисленных коэффициентов с эталонными
Fig. 2. Comparison of calculated coefficients with reference ones

Если среднее значение Евклидова расстояния (avdist) меньше порогового (threshold), то аутентификация считается успешной (рис. 3). В противном случае, пользователь не может считаться аутентифицированным.

```
D:/Голосовая аутентификация/test/Polina_source_mfcc_coefficients_7.txt
dist = 23.080732750362355   threshold = 27   avdist = 23.615565325780302

D:/Голосовая аутентификация/test/Polina_source_mfcc_coefficients_8.txt
dist = 23.29284167229744   threshold = 27   avdist = 23.579707142059984

D:/Голосовая аутентификация/test/Polina_source_mfcc_coefficients_9.txt
dist = 23.00362838136269   threshold = 27   avdist = 23.522099265990256

You are authorized!
```

Рис. 3. Пример успешной аутентификации
Fig. 3. Example of successful authentication

Вывод. В ходе выполнения работы был создан модуль аутентификации пользователей по их голосу с помощью вычисления мел-кепстральных коэффициентов. Данный метод биометрической аутентификации является недорогим и относительно простым способом проверки подлинности пользователей.

Несмотря на очевидные преимущества мел-кепстральных коэффициентов, данный метод имеет определённые недостатки. В частности, данные коэффициенты могут быть вычислены неверно, если диктор находится в обстановке со сторонними шумами. Помимо этого, данные коэффициенты не учитывают эмоциональное состояние говорящего. Для устранения данных недостатков могут быть использованы различные частотные фильтры, а также сторонние алгоритмы анализа аудиозаписей.

Библиографический список:

1. Jain A., Hong L., Pankanti S. Biometric identification //Communications of the ACM. – 2000. – Т. 43. – №. 2. – С. 90-98.
2. Размер рынка голосовой биометрии и анализ доли - Отчет об отраслевых исследованиях - Тенденции роста. URL: https://www.mordorintelligence.com/ru/industry-reports/voice-biometrics-market (дата обращения: 11.12.2023).
3. Руководство по биометрии / Р. М. Болл, Дж. Х. Коннел, Ш. Панканти, Н. К. Ратха, Э. У. Сеньор. М.: Техносфера, 2007. 368 с.
4. Stevens S. S., Volkmann J., Newman E. B. A scale for the measurement of the psychological magnitude pitch // The journal of the acoustical society of America. 1937. Т. 8, №. 3. P. 185-190.
5. Гоноровский И. С. Радиотехнические цепи и сигналы: учебник для вузов. М.: «Сов. радио», 1986. 512 с.
6. Судьенкова А. В. Обзор методов извлечения акустических признаков речи в задаче распознавания диктора // Сборник научных трудов Новосибирского государственного технического университета. 2019. №. 3-4. С. 139-164.

7. Alim S. A., Rashid N. K. A. Some commonly used speech feature extraction algorithms // IntechOpen. 2018. P. 2-19.
8. Allen J. B., Rabiner L. R. A unified approach to short-time Fourier analysis and synthesis // Proceedings of the IEEE. 1977. T. 65, №. 11. P. 1558-1564.
9. Charan R., Manisha A., Karthik R., Rajesh K. M. A text-independent speaker verification model: A comparative analysis // 2017 International Conference on Intelligent Computing and Control (I2C2). – IEEE. 2017. P. 1-6.
10. Misra S. et al. Comparison of MFCC and LPCC for a fixed phrase speaker verification system, time complexity and failure analysis //2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015]. – IEEE, 2015. – С. 1-4.
11. SpeechRecognition-3.8.1. URL: https://pypi.org/project/SpeechRecognition/ (дата обращения: 10.08.2023).
12. Swaroop C. H. A byte of python. – Independent, 2013.
13. Pythonnet – Python.NET URL: https://github.com/pythonnet/pythonnet (дата обращения: 10.08.2023).
14. PyAudio-0.2.11. URL: https://pypi.org/project/PyAudio/ (дата обращения: 10.08.2023).
15. Random-Word-1.0.7. URL: https://pypi.org/project/Random-Word/ (дата обращения: 10.08.2023).

References:

1. Jain A., Hong L., Pankanti S. Biometric identification. Communications of the ACM. 2000; 43(2): 90-98.
2. Voice Biometrics market size and share analysis - Industry Research Report - Growth Trends. URL: https://www.mordorintelligence.com/ru/industry-reports/voice-biometrics-market (accessed date: 11.12.2023). (In Russ)
3. Biometrics Manual / R. M. Ball, J. H. Connell, S. Pancanti, N. K. Ratha, E. U. Senior. M.: Technosphere, 2007; 368 .
4. Stevens S. S., Volkmann J., Newman E. B. A scale for the measurement of the psychological magnitude pitch. *The journal of the acoustic society of America*. 1937; 8(3):185-190.
5. Gonorovskiy I. S. Radio engineering circuits and signals: textbook for universities. *M.: "Soviet radio",* 1986;512. (In Russ)
6. Sudyenkova A.V. Review of methods for extracting acoustic signs of speech in the speaker recognition problem. *Collection of scientific papers of Novosibirsk State Technical University*. 2019; 3-4:139-164. (In Russ)
7. Alim S. A., Rashid N. K. A. Some commonly used speech feature extraction algorithms. *IntechOpen*. 2018; 2-19.
8. Allen J. B., Rabiner L. R. A unified approach to short-time Fourier analysis and synthesis. *Proceedings of the IEEE.* 1977; 65(11):1558-1564.
9. Charan R., Manisha A., Karthik R., Rajesh K. M. A text-independent speaker verification model: A comparative analysis. 2017 International Conference on Intelligent Computing and Control (I2C2). IEEE. 2017; 1-6.
10. Misra S. et al. Comparison of MFCC and LPCC for a fixed phrase speaker verification system, time complexity and failure analysis. 2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015]. IEEE, 2015; 1-4.
11. SpeechRecognition-3.8.1. URL: https://pypi.org/project/SpeechRecognition / (accessed date: 10.08.2023).
12. Swaroop C. H. A byte of python. – Independent, 2013.
13. Pythonnet – Python.NET URL: https://github.com/pythonnet/pythonnet (accessed date: 10.08.2023).
14. PyAudio-0.2.11. URL: https://pypi.org/project/PyAudio / (accessed date: 10.08.2023).
15. Random-Word-1.0.7. URL: https://pypi.org/project/Random-Word / (accessed date: 10.08.2023).

Сведения об авторах:

Елизаров Дмитрий Александрович, кандидат технических наук, доцент, доцент, кафедра «Информационная безопасность»; elizarovdaib@gmail.com

Ашаева Полина Александровна, аспирант, кафедра «Информационная безопасность»; elizarovdaib@gmail.com

Степанова Елизавета Андреевна, кандидат технических наук, доцент, кафедра «Информационная безопасность»; elizarovdaib@gmail.com

Information about the authors:

Dmitriy A. Elizarov, Cand. Sci. (Eng.), Assoc. Prof., Department of Information security; elizarovdaib@gmail.com

Polina A. Ashaeva, Postgraduate Student, Department «Information security»; elizarovdaib@gmail.com

Elizaveta A. Stepanova, Cand. Sci. (Eng.), Assoc. Prof., Department «Information security»; elizarovdaib@gmail.com

Конфликт интересов/Conflict of interest.

Авторы заявляют об отсутствии конфликта интересов/The authors declare no conflict of interest.

Поступила в редакцию/ Received 11.01.2024.

Одобрена после рецензирования/ Reviced 20.02.2024.

Принята в печать/ Accepted for publication 20.02.2024.